BeNeLux Bioinformatics Conference – Brussels, December 2013
Abstract ID: **000** Category: Abstract template

TRANSPOSABLE ELEMENT ANNOTATION USING RELATIONAL RANDOM FORESTS

Eduardo P Costa[1], Leander Schietgat[1,], Ricardo Cerri[2], Celine Vens[1,3,4], Carlos N Fischer[5], Claudia M A Carareto[6], Jan Ramon[1] & Hendrik Blockeel[1,7].*

Department of Computer Science, KU Leuven[1]; Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil[2]; VIB Inflammation Research Center, Ghent[3]; Department of Respiratory Medicine, Ghent University[4]; Department of Statistics, Applied Mathematics, and Computer Science[5], and Department of Biology[6], UNESP São Paulo State University, Brazil; Leiden Institute of Advanced Computer Science[7]. [*]Leander.Schietgat@cs.kuleuven.be

Transposable elements (TEs) are DNA sequences that can change their location within the genome. They contribute to genetic diversity within and across species and their transposing mechanisms may also affect the functionality of genes. Accurate annotation of TEs is an important step towards understanding their effects on genes and their role in genome evolution. We present a framework for annotating TEs which is based on relational random forests. It allows to naturally represent the structured data and biological processes involving TEs. Furthermore, it allows the integration of background knowledge.

INTRODUCTION

Currently, the annotation of TEs involves a fair amount of manual labor. Automated methods exist that screen DNA for candidate TEs, but human annotators take over from there. In this work, we explore how inductive logic programming (ILP) can be used to improve the screening. The framework we propose uses existing methods to create a logic-based representation for each sequence, and then applies an ILP model. In this work, we focus on predicting LTR retrotransposons, a particular type of TEs that is characterized by having long terminal repeats (LTRs) at the boundaries.

METHODS

We propose the following three-step framework [1].

1. The genome is screened for potential LTR retrotransposons. To that aim, we use the tool LTR Finder [2], which scans a DNA sequence to search for matching string pairs (the LTRs), and then filters the list by checking user defined length restrictions. Each remaining candidate, i.e., the region bounded by the LTR pairs, receives a score, depending on how many of a predefined set of structural elements are found in there. The output of this first step is a list of candidate LTR retrotransposons, to be further filtered.
2. Every candidate TE sequence, obtained in the previous step, is screened for the occurrence of protein domains that are known to occur in LTR retrotransposons. Domains are recognized using a profile hidden Markov model (HMM) trained on a multiple sequence alignment corresponding to that subdomain.
3. Each candidate sequence is represented in a first order logic format, by simply listing all its predicted protein domains, and the location in the sequence where that domain was found (see Figure 1). For a given sequence, this representation is fed into an ILP model, together with biological background knowledge. The model predicts for every LTR retrotransposon superfamily the probability that the sequence belongs to that family.

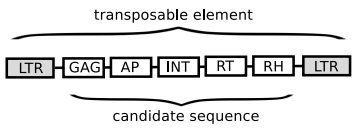

***FIGURE 1**. (a) Typical structure of a TE, delimited by LTRs and annotated with protein domains. (b) Example of an interpretation, consisting of protein domain predictions. For each domain prediction, we have the candidate ID, the domain, the start and end positions in the sequence, and the e-value for the HMM prediction.*

For the ILP model, which is to be learned from data, the learning process is as follows. For each LTR retrotransposon superfamily, a separate model is learned that maps a sequence, represented as above, to the probability that the sequence belongs to that superfamily. This model is built using the FORF approach (first-order random forests) [3]. The language bias includes the following types of tests that are allowed in the nodes of the trees: (1) the occurrence of a particular protein domain, (2) the occurrence of a particular protein domain before another domain, and (3) the number of occurrences of a particular protein domain. As domains may have subtypes, we give the hierarchical "is a subtype" relationship as background knowledge.

RESULTS

Preliminary results based on precision-recall analysis show a significant improvement over state-of-the-art techniques.

REFERENCES

1. E. Costa, L. Schietgat, R. Cerri, C. Vens, C. Fischer, C. Carareto, J. Ramon, and H. Blockeel, Annotating transposable elements in the genome using relational decision tree ensembles, 23rd International Conference on Inductive Logic Programming (2013)
2. Xu, Z., Wang, H.: LTR FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research 35(suppl 2) (2007)
3. Van Assche, A., Vens, C., Blockeel, H., Dzeroski, S.: First order random forests: Learning relational classifiers with complex aggregates. Machine Learning 64(1-3) (2006)